OPEN ACCESS

KING'S College LONDON

King's Research Portal

DOI:
10.1126/science.aag0299

Document Version
Peer reviewed version

Link to publication record in King's Research Portal

Citation for published version (APA):
Alexandrov, L. B., Ju, Y. S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., Campbell, P. J., Vineis, P., Phillips, D. H., & Stratton, M. R. (2016). Mutational signatures associated with tobacco smoking in human cancer. *Science*, *354*(6312), 618-622. https://doi.org/10.1126/science.aag0299

Citing this paper
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
•You may not further distribute the material or use it for any profit-making activity or commercial gain
•You may freely distribute the URL identifying the publication in the Research Portal

Take down policy
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. May. 2023

Mutational signatures associated with tobacco smoking in human cancer

Ludmil B. Alexandrov[1,2,3,*], Young Seok Ju[4], Kerstin Haase[5], Peter Van Loo[5,6], Iñigo Martincorena[7], Serena Nik-Zainal[7,8], Yasushi Totoki[9], Akihiro Fujimoto[10,11], Hidewaki Nakagawa[10], Tatsuhiro Shibata[9,12], Peter J. Campbell[7,13], Paolo Vineis[14,15], David H. Phillips[16], Michael R. Stratton[7,*]

Affiliations:

[1]Theoretical Biology and Biophysics (T-6), Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[2]Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[3]University of New Mexico Comprehensive Cancer Center, Albuquerque, NM 87102, USA
[4]Graduate School of Medical Science and Engineering (GSMSE), Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea
[5]The Francis Crick Institute, Lincoln's Inn Fields Laboratory, 44 Lincoln's Inn Fields, London WC2A 3LY
[6]Human Genome Laboratory, Department of Human Genetics, University of Leuven, 3000 Leuven, Belgium
[7]Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, Cambridgeshire, UK
[8]Department of Medical Genetics, Addenbrooke's Hospital National Health Service (NHS) Trust, Cambridge, UK
[9]Division of Cancer Genomics, National Cancer Center Research Institute, Chuo-ku, Tokyo, Japan.
[10]Laboratory for Genome Sequencing Analysis, RIKEN Center for Integrative Medical Sciences, Tokyo, Japan
[11]Department of Drug Discovery Medicine, Kyoto University Graduate School of Medicine, Kyoto, 606-8507, Japan
[12]Laboratory of Molecular Medicine, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Minato-ku, Tokyo, Japan
[13]Department of Haematology, University of Cambridge, Cambridge CB2 0XY, UK
[14]Human Genetics Foundation (HuGeF), 10126 Torino, Italy
[15]Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, Norfolk Place, London W2 1PG, UK
[16]King's College London, MRC-PHE Centre for Environment and Health, Analytical and Environmental Sciences Division, Franklin-Wilkins Building, 150 Stamford Street, London SE1 9NH, UK

*Correspondence should be addressed to lba@lanl.gov and mrs@sanger.ac.uk.

ABSTRACT

Tobacco smoking increases the risk of at least 17 classes of cancer. We analyzed somatic mutations and DNA methylation in 5,243 cancers of types for which tobacco smoking confers an elevated risk. Smoking is associated with increased mutation burdens of multiple distinct mutational signatures, which contribute to different extents in different cancers. One of these signatures, mainly found in cancers derived from tissues directly exposed to tobacco smoke, is attributable to misreplication of DNA damage caused by tobacco carcinogens. Others likely reflect indirect activation of DNA editing by APOBEC cytidine deaminases and of an endogenous clock-like mutational process. Smoking is associated with limited differences in methylation. The results are consistent with the proposition that smoking increases cancer risk by increasing the somatic mutation load, although direct evidence for this mechanism is lacking in some smoking-related cancer types.

ONE SENTENCE SUMMARY: Multiple distinct mutational processes associated with tobacco smoking in cancer reflect direct and indirect effects of tobacco smoke.

MAIN TEXT

Tobacco smoking has been associated with at least 17 types of human cancer (Table 1) and claims the lives of more than six million people every year (*1-4*). Tobacco smoke is a complex mixture of chemicals among which at least 60 are carcinogens (*5*). Many of these are thought to cause cancer by inducing DNA damage which, if misreplicated, leads to an increased burden of somatic mutations and hence an elevated chance of acquiring "driver" mutations in cancer genes. Such damage is often in the form of covalent bonding of metabolically activated reactive species of the carcinogen to DNA bases, termed "DNA adducts" (*6*). Tissues directly exposed to tobacco smoke (*e.g.* lung) as well as some tissues not directly exposed (*e.g.* bladder) show elevated levels of DNA adducts in smokers and thus evidence of exposure to carcinogenic components of tobacco smoke (*7, 8*).

Each biological process causing mutations in somatic cells leaves a mutational signature (*9*). Many cancers have a somatic mutation in the *TP53* gene and catalogues of *TP53* mutations compiled two decades ago enabled early exploration of these signatures (*10*) showing that lung cancers from smokers have more C>A transversions than lung cancers from non-smokers (*11-14*). To investigate mutational signatures using the thousands of mutation catalogues generated by systematic cancer genome sequencing, we recently described a framework in which each base substitution signature is characterized using a 96 mutation classification that includes the six substitution types together with the bases immediately 5' and 3' to the mutated base (*15*). The analysis extracts mutational signatures from mutation catalogues and estimates the number of mutations contributed by each signature to each cancer genome (*15*). Using this approach, more than 30 different base substitution signatures have been identified (*16-18*).

Here, we studied 5,243 cancer genome sequences (4,633 exomes and 610 whole genomes) of cancer classes for which smoking increases risk to identify mutational signatures and methylation changes associated with tobacco smoking (Table S1). 2,490 samples were reported to be from tobacco smokers and 1,063 from never smokers (Table 1) enabling investigation of the mutational

consequences of smoking by comparing somatic mutations and methylation in smokers with non-smokers for lung, larynx, pharynx, oral cavity, esophagus, bladder, liver, cervix, kidney and pancreas cancers (Fig. 1 and Table S2).

We first compared total numbers of base substitutions, small insertions and deletions (indels) and genomic rearrangements. Total base substitutions were higher in smokers compared to non-smokers for all cancer types together (q-value<0.05) and, for individual cancer types, in lung adenocarcinoma, larynx, liver and kidney cancers (Table S2). Total numbers of indels were higher in smokers compared to non-smokers in lung adenocarcinoma and liver cancer (Table S2). The whole genome sequenced cases allowed comparison of genome rearrangements between smokers and non-smokers in pancreatic and liver cancer, where no differences were found (Table S2). However, sub-chromosomal copy number changes entail genomic rearrangement and can serve as surrogates for rearrangements. Lung adenocarcinomas from smokers exhibited more copy number aberrations than from non-smokers (Table S2).

We then extracted mutational signatures, estimated the contributions of each signature to each cancer and compared the numbers of mutations attributable to each signature in smokers and non-smokers. Increases in smokers compared to non-smokers were seen for signatures 2, 4, 5, 13 and 16 (the mutational signature nomenclature is that used in COSMIC and references (*16-18*)). There was sufficient statistical power to show that these increases were of clonal mutations (mutations present in all cells of each cancer) for signatures 4 and 5 (q-value<0.05) as expected if they are due to cigarette smoke exposure prior to neoplastic change (Supplementary Text).

Signature 4 is characterized mainly by C>A mutations with smaller contributions from other base substitution classes (Fig. 2B and Fig. S1). It was found only in cancer types in which tobacco smoking increases risk and mainly in those derived from epithelia directly exposed to tobacco smoke (Fig. S2 and S3). Signature 4 is very similar to the mutational signature induced *in vitro* by exposing cells to benzo[*a*]pyrene (cosine similarity=0.94; Figs. 2B and S3), a tobacco smoke carcinogen (*19*). The

similarity extends to the presence of a transcriptional strand bias indicative of transcription-coupled nucleotide excision repair (NER) of bulky DNA adducts on guanine (Fig. S1), the proposed mechanism of DNA damage by benzo[*a*]pyrene. Thus, signature 4 is likely the direct mutational consequence of misreplication of DNA damage induced by tobacco carcinogens.

Most lung and larynx cancers from smokers had many signature 4 mutations. There were more signature 4 mutations in cancers from smokers compared to non-smokers in all cancer types together (Table S2) and in lung squamous, lung adenocarcinoma and larynx cancers (Table S2) accounting, in large part, for differences in total numbers of base substitutions (Table 1). 13.8% of lung cancers in non-smokers showed many signature 4 mutations (Fig. 2A;>1 mutation per MB) which may be due to passive smoking, misreporting of smoking habits or annotation errors. Signature 4 mutations were also found in oral cavity, pharynx and esophagus cancers, albeit in much smaller numbers than in lung and larynx cancers perhaps due to less exposure to tobacco smoke or more efficient clearance. Differences in mutation burden attributed to signature 4 between smokers and non-smokers were not observed in these cancer types (Fig. 1). Signature 4 mutations were found at low levels in cancers of the liver, an organ not directly exposed to tobacco smoke, and were elevated in smokers compared to non-smokers (Fig. 1).

Signature 4 was not extracted from bladder, cervix, kidney or pancreas cancers, despite the known risks conferred by smoking and the presence of many smokers in these series. It was also not extracted from cancers of the stomach, colorectum, ovary and acute myeloid leukemia for which the smoking status in the analyzed series was unknown but among which many are likely to have been smokers. The tissues from which all these cancer types are derived are not directly exposed to tobacco smoke. Simulations indicate that the lack of signature 4 is not due to statistical limitations (Supplementary Text and Fig. S4). The absence of signature 4 suggests that misreplication of direct DNA damage due to tobacco smoke constituents does not contribute substantially to mutation

burden in these cancers even though DNA adducts indicative of tobacco-induced DNA damage are present in the tissues from which they arise (*7*).

Signatures 2 and 13 are characterized by C>T and C>G mutations respectively at TpC dinucleotides and have been attributed to overactive DNA editing by APOBEC deaminases (*20*). The cause of the over-activity in most cancers has not been established although APOBECs are implicated in the cellular response to entrance of foreign DNA, retrotransposon movement and local inflammation (*21*). Signatures 2 and 13 showed more mutations in smokers than non-smokers in lung adenocarcinoma (Table S2). Since they are found in many other cancer types, where they are apparently unrelated to tobacco smoking, it seems unlikely that the signature 2 and 13 mutations associated with smoking in lung adenocarcinoma are direct consequences of misreplication of DNA damage induced by tobacco smoke. More plausibly, the cellular machinery underlying signatures 2 and 13 is activated by tobacco smoke, perhaps as a result of inflammation arising from deposition of particulate matter or by indirect consequences of DNA damage.

Signature 5 is characterized by mutations distributed across all 96 subtypes of base substitution, with predominance of T>C and C>T mutations (Fig. 2B) and evidence of transcriptional strand bias for T>C mutations (*18*). Signature 5 is found in all cancer types, including those unrelated to tobacco smoking, and in most cancer samples. It is "clock-like" in that the number of mutations attributable to this signature correlates with age of diagnosis in many cancer types (*17*). Signature 5, together with signature 1, is thought to contribute to mutation accumulation in most normal somatic cells and in the germline (*17, 22*). The mechanisms underlying signature 5 are not well understood, although an enrichment of signature 5 mutations was found in bladder cancers harboring inactivating mutations in *ERCC2* which encodes a component of NER (*23*).

Signature 5 (or a similar signature that is difficult to differentiate from it because of their relatively flat profiles) was increased between 1.3-fold and 5.1-fold (q-value<0.05; Table S2) in smokers compared to non-smokers in all cancer types together and in lung squamous, lung adenocarcinoma,

larynx, pharynx, oral cavity, esophageal squamous, bladder, liver and kidney cancers. The association of smoking with signature 5 mutations across these nine cancer types therefore includes some for which the risks conferred by smoking are modest and for which normal progenitor cells are not directly exposed to cigarette smoke (Table 1). Given the clock-like nature of signature 5 (*17*), its presence in the human germline (*22*), its ubiquity in cancer types unrelated to tobacco smoking (*18*) and its widespread occurrence in non-smokers, it seems unlikely that signature 5 mutations associated with tobacco smoking are direct consequences of misreplication of DNA damaged by tobacco carcinogens. It is more plausible that smoking affects the machinery generating signature 5 mutations (*23*). Presumably as a consequence of the effects of smoking, signature 5 mutations correlated with age of diagnosis in non-smokers (p-value:0.001) but not in smokers (p-value:0.59).

Signature 16 is predominantly characterized by T>C mutations at ApT dinucleotides (Fig. 2B), exhibits a strong transcriptional strand bias consistent with almost all damage occurring on adenine (Fig. S5) and has only been found thus far in liver cancer. The underlying mutational process is currently unknown. Signature 16 exhibited a higher mutation burden in smokers compared to non-smokers in liver cancer (Table S2).

For smokers with lung, larynx, pharynx, oral cavity, esophageal, bladder, liver, cervix, kidney and pancreas cancers quantitative data on cumulative exposure to tobacco smoke were available (Table S1). Total numbers of base substitution mutations positively correlated with pack years smoked for all cancer types together (q-value<0.05) and for lung adenocarcinoma (Table S3). For individual mutational signatures, correlations with pack years smoked were found in multiple cancer types for signatures 4 and 5 (Table S3). Signature 4 correlated with pack years in lung squamous, lung adenocarcinoma, larynx and liver cancers. Signature 5 correlated with pack years in all cancers together, in lung adenocarcinoma, pharynx, oral cavity and bladder cancers (Table S3). In lung adenocarcinoma, correlations with pack years smoked were also observed for signatures 2 and 13. The rates of these correlations allow estimation of the approximate numbers of mutations

accumulated in a normal cell of each tissue due to smoking a pack of cigarettes a day for a year: lung, 150 mutations; larynx, 97; pharynx, 39; oral cavity 23; bladder, 18; liver, 6 (Table S3).

Consistent with our results, previous studies have reported the higher numbers in smokers compared to non-smokers of total base substitutions in lung adenocarcinoma (mainly due to C>A substitutions) (*24, 25*), of signatures 4 and 5 in lung adenocarcinoma (*18*), of signature 4 in liver cancer (*26*) and of signature 5 in bladder cancer (*23*).

Differential methylation of the DNA of normal cells of smokers compared to non-smokers has been reported (*27*). Using data from methylation arrays, each containing ~470,000 of the ~28 million CpG sites in the human genome, we evaluated whether differences in methylation are found in cancers. Overall levels of CpG methylation in DNA from cancers were similar in smokers and non-smokers for all cancer types (Fig. S6). Individual CpGs were differentially methylated (>5% difference) only in two cancer types: 369 CpGs were hypo- and 65 were hyper-methylated in lung adenocarcinoma, with 5 hypo- and 3 hyper-methylated in oral cancer (Fig. 3 and S7). CpGs exhibiting differences in methylation clustered in certain genes but were neither associated with known cancer genes more than expected by chance nor with genes hypo-methylated in normal blood or buccal cells of tobacco smokers (Fig. S8; Tables S4 and S5) (*27*). Therefore, with the exception of lung cancer, CpG methylation showed limited differences between the cancers of smokers and non-smokers (Fig. 3).

The genomes of smoking-associated cancers permit reassessment of our understanding of how tobacco smoke causes cancer. Consistent with the proposition that an increased mutation load caused by tobacco smoke contributes to increased cancer risk, the total mutation burden is elevated in smokers compared to non-smokers in lung adenocarcinoma, larynx, liver and kidney cancers. However, differences in total mutation burden were not observed in the other smoking-associated cancer types and in some there were no statistically significant smoking-associated differences in mutation load, signatures or DNA methylation. Caution should be exercised in the interpretation of the latter observations. In addition to limitations of statistical power, multiple rounds of clonal

expansion over many years are often required for development of a symptomatic cancer. It is thus conceivable that, in the normal tissues from which smoking associated cancer types originate, there are more somatic mutations in smokers than in non-smokers (or differences in methylation) but that these differences become obscured during the intervening clonal evolution. Moreover, some theoretical models predict that relatively small differences in mutation burden caused by smoking in pre-neoplastic cells could account for the observed increases in cancer risks (*28*) and others that differences in mutation burden between smokers and non-smokers need not be observed in the final cancers (Supplementary Text and Fig. S6). Thus, increased somatic mutation loads in precancerous tissues may still explain the smoking-induced risks of most cancers, although other mechanisms have been proposed (*29, 30*).

The generation of the increased somatic mutation burden by tobacco smoking, however, appears to be mechanistically complex. Smoking correlates with increases in base substitutions of multiple mutational signatures, together with increases in indels and copy number changes. The extent to which these distinct mutational processes operate differs between tissue types, at least in part depending on the degree of direct exposure to tobacco smoke, and their mechanisms range from misreplication of DNA damage caused by tobacco smoke constituents to activation of more generally operative mutational processes. Although we cannot exclude roles for covariate behaviours of smokers or differences in the biology of cancers arising in smokers compared to non-smokers, smoking itself is most plausibly the cause of these differences.

REFERENCES AND NOTES

1. B. Secretan *et al.*, A review of human carcinogens--Part E: tobacco, areca nut, alcohol, coal smoke, and salted fish. *Lancet Oncol* **10**, 1033-1034 (2009).
2. S. S. Lim *et al.*, A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2224-2260 (2012).
3. B. Pesch *et al.*, Cigarette smoking and lung cancer--relative risk estimates for the major histological types from a pooled analysis of case-control studies. *International journal of cancer* **131**, 1210-1219 (2012).
4. A. Agudo *et al.*, Impact of cigarette smoking on cancer risk in the European prospective investigation into cancer and nutrition study. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **30**, 4550-4557 (2012).
5. S. S. Hecht, Tobacco carcinogens, their biomarkers and tobacco-induced cancer. *Nature reviews. Cancer* **3**, 733-744 (2003).
6. D. H. Phillips, The formation of DNA adducts. In *The Cancer Handbook,* 2nd edition, M. R. Allison, Ed. (Wiley, Chichester, 2007), pp. 338-350.
7. D. H. Phillips, Smoking-related DNA and protein adducts in human tissues. *Carcinogenesis* **23**, 1979-2004 (2002).
8. D. H. Phillips, S. Venitt, DNA and protein adducts in human tissues resulting from exposure to tobacco smoke. *International journal of cancer* **131**, 2733-2753 (2012).
9. L. B. Alexandrov, M. R. Stratton, Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Current opinion in genetics & development* **24**, 52-60 (2014).
10. P. Hainaut, M. Hollstein, p53 and human cancer: the first ten thousand mutations. *Adv. Cancer Res.* **77**, 81-137 (2000).
11. M. F. Denissenko, A. Pao, M. Tang, G. P. Pfeifer, Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in P53. *Science* **274**, 430-432 (1996).
12. G. P. Pfeifer, M. F. Denissenko, Formation and repair of DNA lesions in the p53 gene: relation to cancer mutations? *Environ Mol Mutagen* **31**, 197-205 (1998).
13. L. E. Smith *et al.*, Targeting of lung cancer mutational hotspots by polycyclic aromatic hydrocarbons. *J Natl Cancer Inst* **92**, 803-811 (2000).
14. F. Le Calvez *et al.*, TP53 and KRAS mutation load and types in lung cancers in relation to tobacco smoke: distinct patterns in never, former, and current smokers. *Cancer Res* **65**, 5076-5083 (2005).
15. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, M. R. Stratton, Deciphering signatures of mutational processes operative in human cancer. *Cell reports* **3**, 246-259 (2013).
16. L. B. Alexandrov, Understanding the origins of human cancer. *Science* **350**, 1175 (2015).
17. L. B. Alexandrov *et al.*, Clock-like mutational processes in human somatic cells. *Nature genetics* **47**, 1402-1407 (2015).
18. L. B. Alexandrov *et al.*, Signatures of mutational processes in human cancer. *Nature* **500**, 415-421 (2013).
19. S. Nik-Zainal *et al.*, The genome as a record of environmental exposure. *Mutagenesis* **30**, 763-770 (2015).
20. S. Nik-Zainal *et al.*, Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-993 (2012).
21. C. Swanton, N. McGranahan, G. J. Starrett, R. S. Harris, APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity. *Cancer discovery* **5**, 704-712 (2015).
22. R. Rahbari *et al.*, Timing, rates and spectra of human germline mutation. *Nature genetics* **48**, 126-133 (2016).

23. J. Kim *et al.*, Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nature genetics* **advance online publication**, (2016).
24. R. Govindan *et al.*, Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150**, 1121-1134 (2012).
25. M. Imielinski *et al.*, Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107-1120 (2012).
26. A. Fujimoto *et al.*, Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nature genetics* **48**, 500-509 (2016).
27. A. E. Teschendorff *et al.*, Correlation of Smoking-Associated DNA Methylation Changes in Buccal Cells With DNA Methylation Changes in Epithelial Cancer. *JAMA oncology* **1**, 476-485 (2015).
28. C. Tomasetti, L. Marchionni, M. A. Nowak, G. Parmigiani, B. Vogelstein, Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 118-123 (2015).
29. M. Sopori, Effects of cigarette smoke on the immune system. *Nature reviews. Immunology* **2**, 372-377 (2002).
30. H. Rubin, Selective clonal expansion and microenvironmental permissiveness in tobacco carcinogenesis. *Oncogene* **21**, 7392-7411 (2002).

ACKNOWLEDGEMENTS

This work was supported by the Wellcome Trust (grant number 098051). S.N.-Z. is a Wellcome-Beit Prize Fellow and is supported through a Wellcome Trust Intermediate Fellowship (grant WT100183MA). P.J.C. is personally funded through a Wellcome Trust Senior Clinical Research Fellowship (grant WT088340MA). L.B.A. is personally supported through a J. Robert Oppenheimer Fellowship at Los Alamos National Laboratory. This research used resources provided by the Los Alamos National Laboratory Institutional Computing Program, which is supported by the U.S. Department of Energy National Nuclear Security Administration under Contract No. DE-AC52-06NA25396. Research performed at Los Alamos National Laboratory was carried out under the auspices of the National Nuclear Security Administration of the United States Department of Energy. This work was supported by the Francis Crick Institute which receives its core funding from Cancer Research UK, the UK Medical Research Council, and the Wellcome Trust. D.H.P. is funded by Cancer Research UK (grant number C313/A14329), the Wellcome Trust (Grants 101126/Z/13/Z and 101126/B/13/Z), the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Health Impact of Environmental Hazards at King's College London in partnership with Public Health England (PHE) and by the project EXPOSOMICS, grant agreement 308610-FP7 (European Commission). P.V. was partially supported by the project EXPOSOMICS, grant agreement 308610-FP7 (European Commission). Y.T. and T.S. are supported by the Practical Research for Innovative Cancer Control from Japan Agency for Medical Research and Development (15ck0106094h0002), and National Cancer Center Research and Development Funds (26-A-5). We would like to thank The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC), and the authors of all studies cited in Tables S1 for providing free access to their somatic mutational data.

FIGURE LEGENDS

Fig. 1. Comparison between tobacco smokers and lifelong non-smokers. Bars are used to display average values for numbers of somatic substitutions per megabase, numbers of indels per megabase, numbers of dinucleotide mutations per megabase, numbers of breakpoints per megabase, fraction of the genome that shows copy number changes and numbers of mutations per megabase attributed to mutational signatures found in multiple cancer types associated with tobacco smoking. Light gray bars are non-smokers, while dark gray bars are smokers. Comparisons between smokers and non-smokers for all features, including mutational signatures specific for a cancer type and overall DNA methylation are provided in Table S2. Error bars correspond to 95% confidence intervals for each feature. Each q-value is based on a two-sample Kolmogorov-Smirnov test corrected for multiple hypothesis testing for all features in a cancer type. Cancer types are ordered based on their age adjusted odds ratios for smoking as provided in Table 1. Data for numbers of breakpoints per megabase and fraction of the genome that shows copy number changes were not available for liver cancer and small cell lung cancer. Adeno stands for Adenocarcinoma; Esophag. stands for Esophagus. Note that the presented data include only a few cases (<10) of nonsmokers for lung small cell, lung squamous and cancer of the larynx.

Fig. 2. Mutational signatures associated with tobacco smoking. (***A***) Each panel contains 25 randomly selected cancer genomes (represented by individual bars) from either smokers or non-smokers in a given cancer type. The y-axes reflect numbers of somatic mutations per megabase. Each bar is colored proportionately to the numbers of mutations per megabase attributed to the mutational signatures found in that sample. Naming of mutational signatures is consistent with previous reports (*16-18*). (***B***) Each panel contains the pattern of a mutational signature associated with tobacco smoking. Signatures are depicted using a 96 substitution classification defined by the substitution type and sequence context immediately 5’ and 3’ to the mutated base. Different colors

are used to display different types of substitutions. The percentages of mutations attributed to specific substitution types are on the vertical axes, while the horizontal axes display different types of substitutions. Mutational signatures are depicted based on the trinucleotide frequency of the whole human genome. Signatures 2, 4, 5, 13 and 16 are extracted from cancers associated with tobacco smoking. The signature of benzo[*a*]pyrene is based on *in vitro* experimental data (*19*). Numerical values for these mutational signatures are provided in Table S6.

Fig. 3. Differentially methylated individual CpGs in tobacco smokers across cancers associated with tobacco smoking. Each dot represents an individual CpG. The horizontal axes reflect differences in methylation between lifelong non-smokers and smokers, where positive values correspond to hyper-methylation and negative values to hypo-methylation. The vertical axes depicted levels of statistical significance. Results satisfying a Bonferroni threshold of 10^{-7} (above the red line) are considered statistically significant.

TABLES

Cancer Type	Odds ratios	Non-smokers	Smokers	Total number of mutational signatures found in the cancer type	Signature 4 found in cancer type	Mutational signatures with elevated mutation burden in smokers compared to non-smokers (q-value<0.05)
All Cancer Types		1,062	2,490	26	Y	4, 5*
Small Cell Lung Cancer	111.3	3	145	6	Y	
Lung Squamous	103.5	7	168	8	Y	4*, 5
Lung Adenocarcinoma	21.9	120	558	7	Y	2*, 4*, 5*, 13*
Larynx	13.2	6	117	5	Y	4*, 5
Pharynx	6.6	27	49	5	Y	5*
Oral Cavity	4.2	98	265	5	Y	5*
Esophagus Squamous	3.9	99	193	9	Y	5
Esophagus Adenocarcinoma	3.9	67	175	9	Y	
Bladder	3.8	111	288	5	N	5*
Liver	2.9	157	235	19	Y	4*, 5, 16
Stomach	2.1	472		13	N	N/A
Acute Myeloid Leukemia	2.0	202		2	N	N/A
Ovary	1.9	458		3	N	N/A
Cervix	1.8	94	74	8	N	
Kidney	1.7	154	103	6	N	5
Pancreas	1.6	119	120	11	N	
Colorectal	1.3	559		4	N	N/A

Table 1. Mutational signatures and cancer types associated with tobacco smoking. Information about the age adjusted odds ratios for current male smokers to develop cancer is taken from refs. (*2-4*). Odds ratios for small cell lung cancer, lung squamous and lung adenocarcinoma are for an average daily dose of more than 30 cigarettes. Odds ratios for cervix and ovary are for current female smokers. Detailed information about all mutation types, all mutational signatures and DNA methylation is provided in Table S2. Nomenclature for signature IDs is consistent with the COSMIC website, http://cancer.sanger.ac.uk/cosmic/signatures. The patterns of all mutational signatures with elevated mutation burden in smokers are displayed in Fig. 2B. N/A denotes lack of smoking annotation for a given cancer type. * denotes that a signature correlates with pack years smoked in a cancer type.

LIST OF SUPPLEMENTARY MATERIALS

Methods

Supplementary Text

Fig. S1. Comparison of transcriptional strand bias between *in vitro* exposure to benzo[*a*]pyrene and signature 4 extracted from all tobacco-associated cancer types.

Fig. S2. Extracting the pattern of signature 4 from different cancer types.

Fig. S3. Similarity between extraction of signature 4 across different cancer types and other known mutational signatures.

Fig. S4. Sensitivity for detecting signature 4 in cervix, bladder and renal cancer.

Fig. S5. Strong transcriptional strand bias of signature 16.

Fig. S6. Comparison between lifelong non-smokers and smokers based on overall CpG methylation profiles.

Fig. S7. CpG islands and differentially methylated CpGs in smokers compared to non-smokers.

Fig. S8. Cancer tissue methylation of individual CpGs near genes differentially methylated in blood or buccal cells of smokers.

Fig. S9. Interquartile range of methylation levels in tobacco-associated cancer types.

Fig. S10. Mutation rate variation across normal cells can lead to complex relationships between cancer mutation burden and cancer risk.

References and notes (*31 – 53*)

Author Contributions

Table S1: Detailed information about each examined tobacco-associated cancer sample.

Table S2: Comparison of features of tobacco smokers to the ones of lifelong non-smokers.

Table S3: Relationships between mutational signatures and pack years smoked.

Table S4: Individual CpGs with differential methylation in lung adenocarcinoma.

Table S5: Individual CpGs with differential methylation in oral cancer.

Table S6: Numerical patterns of mutational signatures associated with tobacco smoking.

Fig. 1

Somatic Substitutions | Other Mutations | Mutational Signatures

All Cancers

Lung Small Cell

Lung Squamous

Lung Adeno

Larynx

Pharynx

Oral Cavity

Esophag. Squamous

Esophag. Adeno

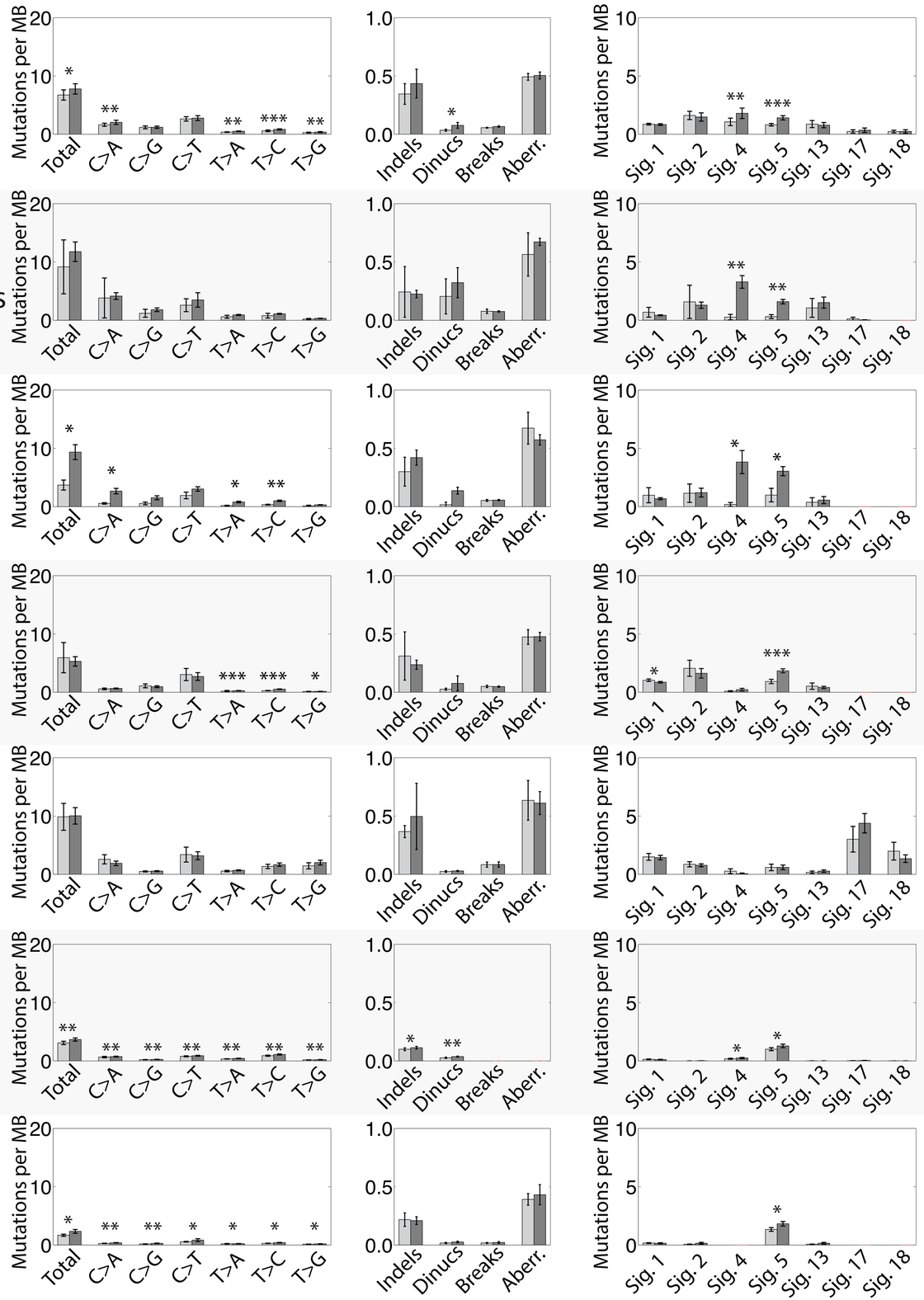
Bladder

Liver

Cervix

Kidney

Pancreas



Lifelong non-smokers Smokers ***q-value < 10^{-5} **q-value < 10^{-2} *q-value < 0.05

A

Tobacco Smokers

Non-Smokers

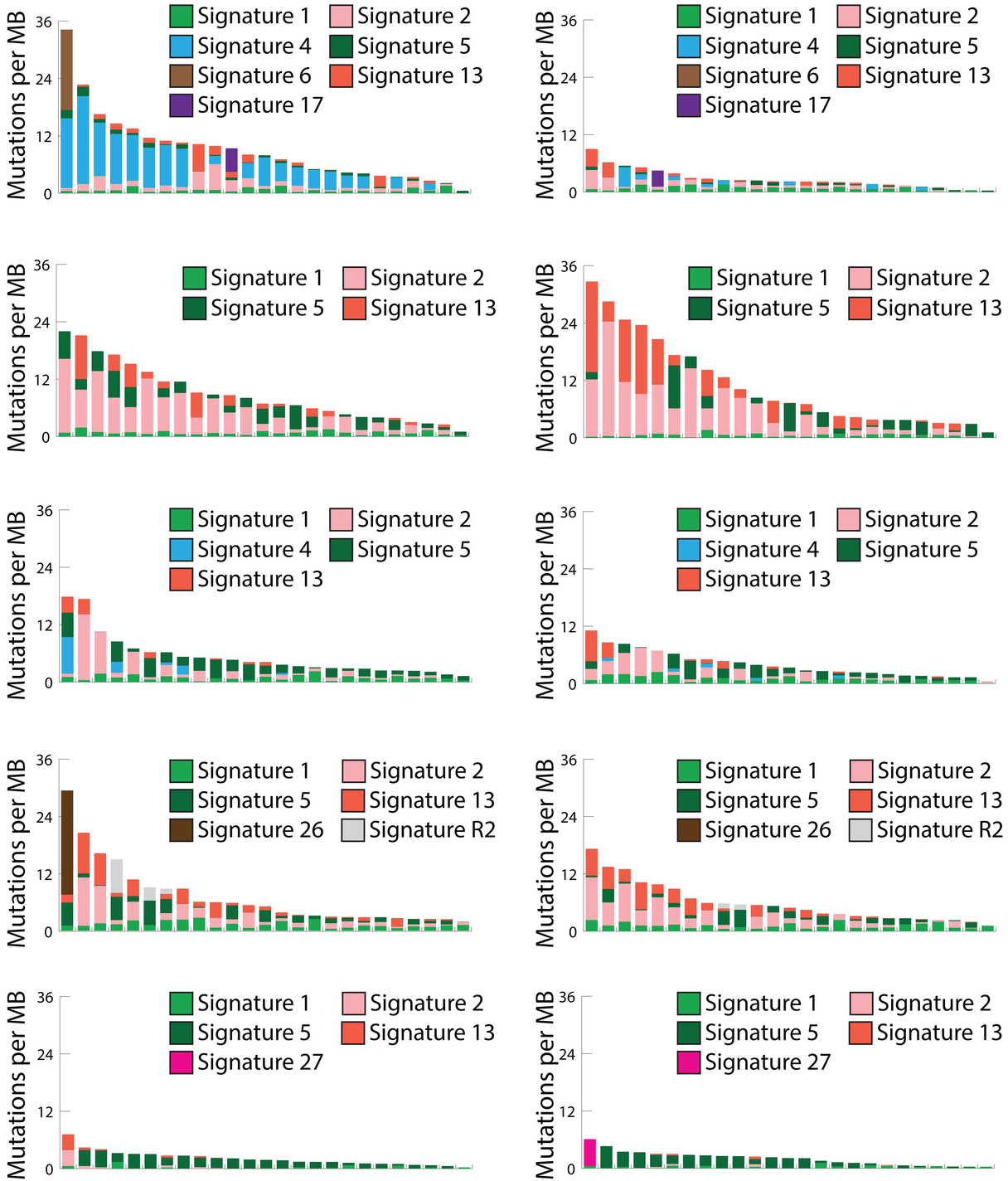
Lung Adeno

Bladder

Oral Cavity

Cervix

Kidney



B

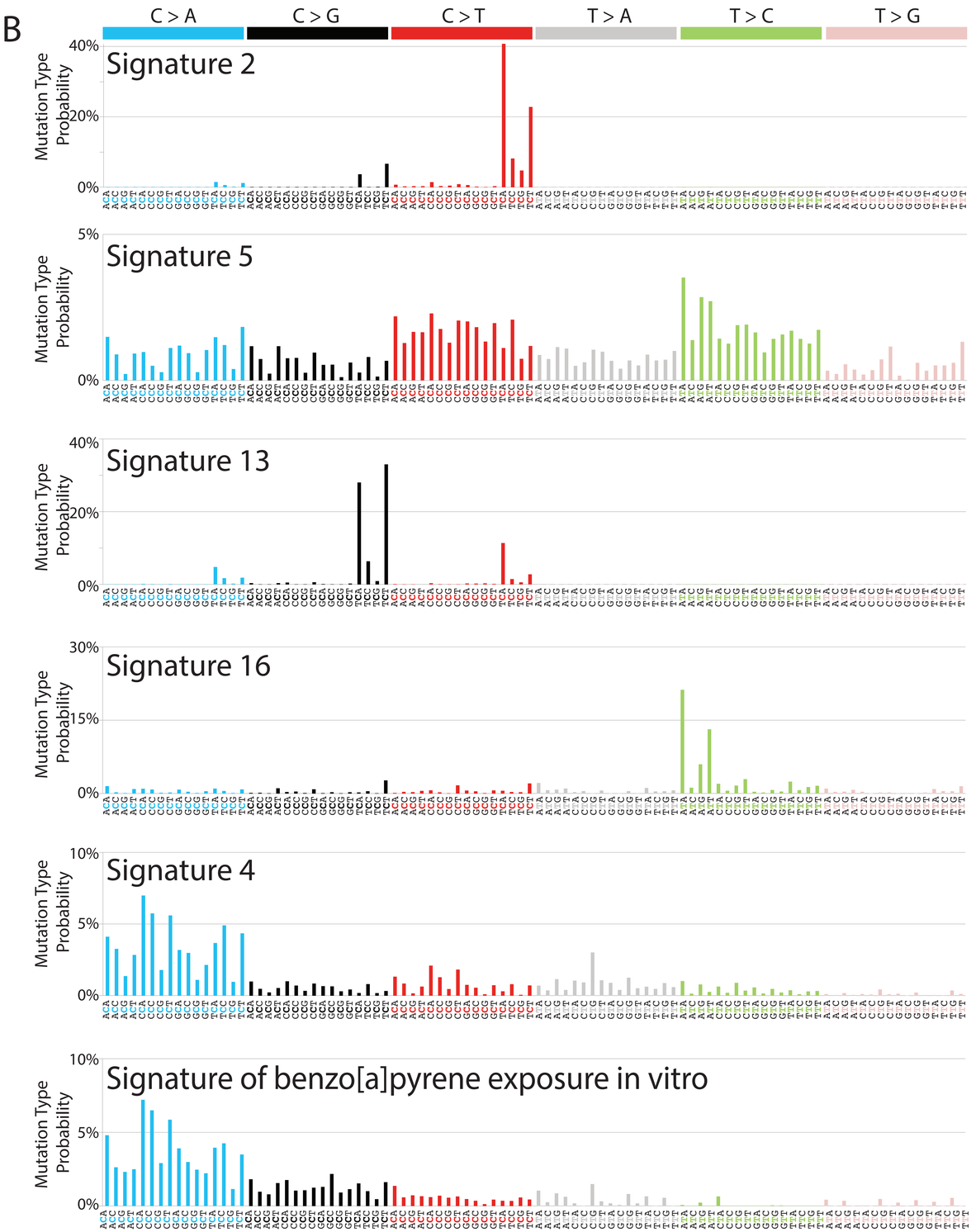


Fig. 2

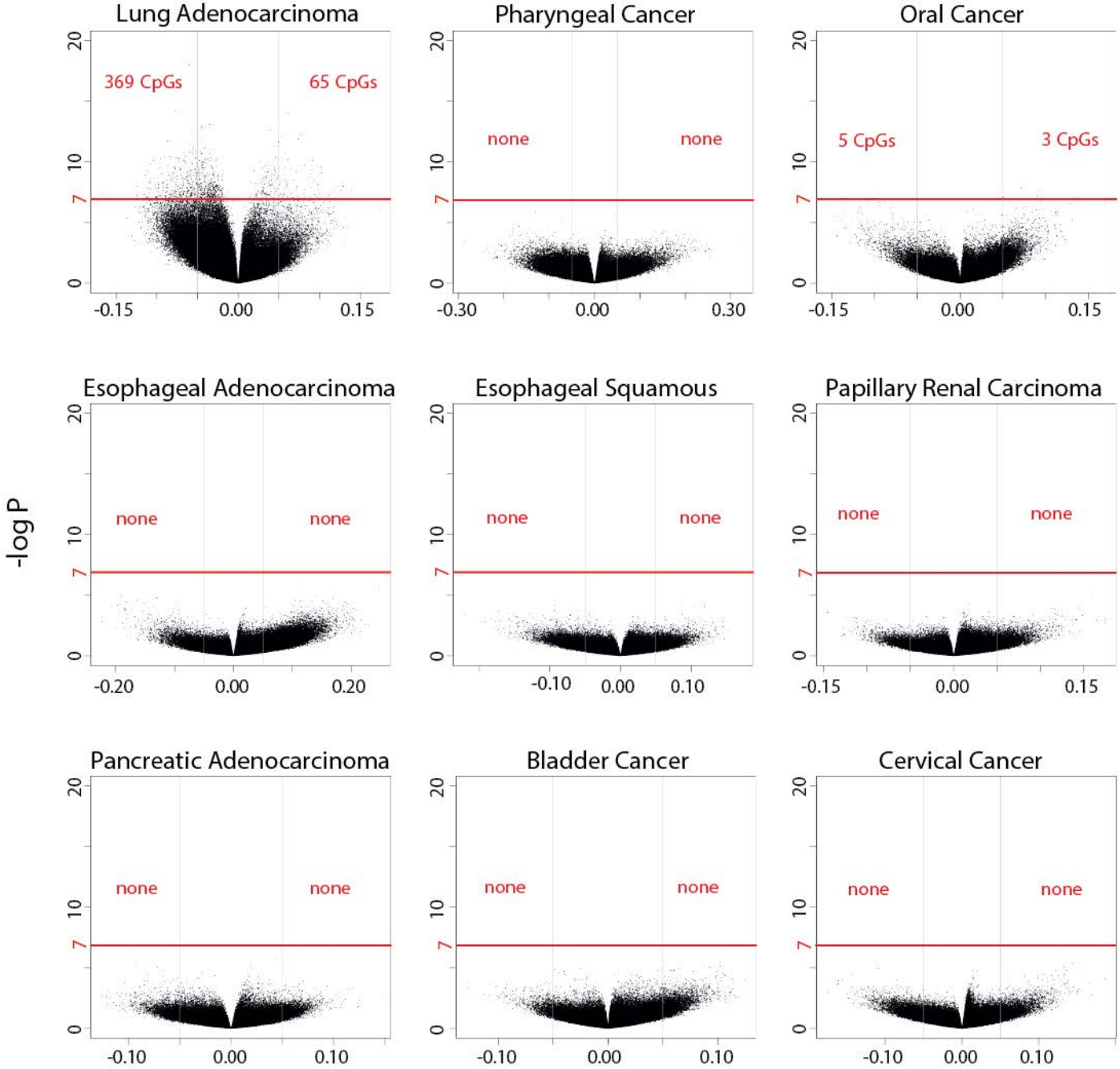

Lung Adenocarcinoma
369 CpGs
65 CpGs
Pharyngeal Cancer
none
none
Oral Cancer
5 CpGs
3 CpGs
Esophageal Adenocarcinoma
none
none
Esophageal Squamous
none
none
Papillary Renal Carcinoma
none
none
Pancreatic Adenocarcinoma
none
none
Bladder Cancer
none
none
Cervical Cancer
none
none
-log P
Average Difference in Methylation
(Smokers - Lifelong non-smokers)

Fig. 3